גישות למידת מכונה לתיארוך כתבי יד עבריים מימי הביניים באמצעות ניתוח קודיקולוגי

אלכסנדר גולדברג, גילה פריבור ואבשלום אלמלח*

מבוא

לכתבי יד עבריים רבים מימי הביניים אין תאריך העתקה ידוע. תיארוך כתבי יד אלה הוא משימה מרכזית במחקר, שכן היעדר תאריך פוגע ביכולת להשתמש בהם כמקורות היסטוריים מהימנים. בדרך כלל, מלאכת התיארוך נתפסת כמיומנות ייחודית, השמורה לקומץ מומחים המסוגלים להעריך את גיל כתב היד על סמך מאפיינים חומריים וסגנוניים.[1]

ידיעת מועד ההעתקה אינה מספקת רק הקשר היסטורי, אלא תורמת להבנת ההשלכות התרבותיות, החברתיות והאינטלקטואליות הגלומות בכתב היד. נוסף על כך, בעזרת תכונות קודיקולוגיות ופליאוגרפיות, תיארוך מדויק יכול לתרום לאימות המסמך. מידע זה מאפשר מחקר בין-תחומי, החוצה גבולות דיסציפלינריים ומגשר בין תובנות מתחומי הארכיאולוגיה, ההיסטוריה ומדעי המידע – ובכך מעמיק את הבנת הטקסט.[2]

התפתחויות טכנולוגיות בתחומי למידת המכונה והזמינות ההולכת וגדלה של מאגרי נתונים המכילים אלפי רשומות של כתבי יד פותחות אפשרויות חדשות לתיארוך אוטומטי. טכנולוגיות של ראייה ממוחשבת ולמידת מכונה מאפשרות ניתוח מתמטי של כתבי יד באמצעות סיווג אוטומטי של מאפיינים קודיקולוגיים ופליאוגרפיים. אוטומציה של תהליכים אלה עשויה לחסוך זמן ומשאבי מחקר ולהנגיש את התיארוך גם מעבר למעגל המומחים המצומצם.[3]

* מר אלכסנדר גולדברג, פרופ׳ גילה פריבור וד״ר אבשלום אלמלח, המחלקה למדעי המידע ויישומי בינה מלאכותית, אוניברסיטת בר-אילן.

1 Sheng He, Petros Sammara, Jan Burgers and Lambert Schomaker *Towards Style-Based Dating of Historical Documents: Paper Presented at the 14th International Conference on Frontiers in Handwriting Recognition*, (Hersonissos, Greece, 2014), pp. 265–270 doi: 10.1109/ICFHR.2014.52.

2 Ahmad Droby, Irina Rabaev, Daria Vasyutinsky Shapira, Berat Kurar Barakat and Jihad El-Sana, "Digital Hebrew Paleography: Script Types and Modes", *Journal of Imaging*, 8(5) (2022): 143. https://doi.org/10.3390/jimaging8050143

3 Daria Vasyutinsky Shapira, Irina Rabaev, Berat Kurar Barakat, Ahmad Droby and Jihad El-Sana, "Deep learning for paleographic analysis of medieval Hebrew manuscripts: a

מחקר זה עוסק בתיארוך של כתבי יד עבריים מימי הביניים באמצעות למידת מכונה מונחית, הנשענת על נתונים קודיקולוגיים, ביבליוגרפיים ותיאוריים המצויים במסדי נתונים כדוגמת "ספרדתא". איכות הנתונים הזמינים משתנה בהתאם לנסיבות: האם החוקר מחזיק בידו את כתב היד הפיזי, או שמא מדובר בעותק דיגיטלי? ואם מדובר בסריקה, האם זו סריקה בצבע או בשחור-לבן ומהי רמת הפירוט שלה? ככל שהגישה אל פרטי כתב היד מלאה ומדויקת יותר, כך משתפרת גם רמת הדיוק של התיארוך האפשרי.

זהו מחקר בינתחומי הפועל בממשק שבין מדעי הרוח למדעי המחשב בתחום המכונה "מדעי הרוח הדיגיטליים". תחום זה עוסק בפיתוח שיטות וכלים ממוחשבים המיועדים להעמקת המחקר ההומניסטי.[4] במסגרת זו המחקר הנוכחי שואף לפתח מודל לחיזוי שנת כתיבתם של כתבי יד עבריים שנוצרו עד שנת 1540 מתוך הסתמכות על המאפיינים הקודיקולוגיים שלהם.

מטרת המחקר להבין מהם המאפיינים המשפיעים ביותר על התאריך, מהו כיוון השפעתם וכיצד הם פועלים במצטבר. אחד מיתרונות השימוש בלמידת מכונה הוא יכולתה לזהות קשרים מורכבים בין מאפיינים שונים, ולכמת את תרומתם לחיזוי שנת הכתיבה. ניתוח של תוצרי המודל מציע לחוקרים תובנות חדשות באשר למאפיינים שהשתנו לאורך התקופות, ולתפקיד שמילאו בהתהוות הדפוסים של הפקת כתבי יד עבריים. בכך, המחקר אינו רק מציע כלי תיארוך אוטומטי, אלא גם פותח אפיק להבנה היסטורית מעמיקה יותר של התמורות שחלו במלאכת כתיבת הספר העברי בימי הביניים ויכול לשמש בסיס למחקרים קודיקולוגיים עתידיים.

סקירת ספרות

כתבי היד העבריים מימי הביניים הם עדויות חומריות לתרבות היהודית לדורותיה, הכתובות באותיות עבריות, ולעיתים גם בערבית יהודית או בשפות אחרות. קולט סיראט מונה שלוש סיבות עיקריות לחשיבות חקרם של כתבי היד העבריים: ראשית, רבים מן הטקסטים המרכזיים בתרבות היהודית נכתבו והועתקו בכתב יד לפני המצאת הדפוס בעברית (שנים 1460–1480 בקירוב), והם השתמרו אך ורק בצורתם הקודמת לדפוס. שנית, שלא כמהדורות הדפוס והדיגיטל הנוטות להציע גרסה אחת של הטקסט, כתבי היד משמרים לעיתים גרסאות שונות. שונות זו פותחת פתח להשוואה טקסטואלית ולהכרעות מדעיות בין נוסחים. שלישית, כתב יד אינו רק כלי להעברת תוכן, אלא גם חפץ היסטורי בעל ערך עצמאי: הוא מספר את

DH team collaboration experience"' *TwinTalks@ DH/DHN* (2020): 84–92. http://ceur-ws.org/Vol-2717/paper09.pdf

4 איתי מרינברג-מיליקובסקי, **מילים שקולות: צעדים ראשונים במחקר הספרות החישובי** (רעננה: למדא – ספרי האוניברסיטה הפתוחה, 2022).

סיפורם של האנשים שהיו מעורבים ביצירתו – המעתיק, הבעלים לדורותיהם ולעיתים אף המחבר – ולעיתים כולל סימנים המעידים על מסעותיו הגאוגרפיים והתרבותיים.[5]

עבודתה של סיראט מדגישה את ההיבט ההיסטורי של חקר כתבי היד ואת תרומתו להבנת הרבדים השונים של התרבות היהודית. מלאכי בית אריה עמד על הפצתם הרחבה של כתבי היד העבריים ברחבי העולם היהודי בתקופת ימי הביניים. הפצה זו תרמה להתפתחותם של סגנונות כתיבה מגוונים וליצירתן של מסורות הפקה ועיצוב ייחודיות לכתבי יד עבריים.[6] בהקשר זה, המחקר הנוכחי מבקש לתרום לביסוס הבנה טובה יותר של הקשרים בין מאפיינים קודיקולוגיים שונים לבין זמן העתקתם של כתבי היד. גילוי דפוסים עקביים בין פרטי העיצוב החומריים לבין טווחי תאריכים עשוי לשפוך אור על השינויים שחלו לאורך הדורות בטכניקות הכתיבה, במנהגי ההפקה ובטעמים האסתטיים של יוצרי כתבי היד.

ספרדתא

ספרדתא הוא מסד נתונים ומערכת מתקדמת לאחזור מידע, המתעדים באופן שיטתי אלפי מאפיינים קודיקולוגיים ונתונים נוספים מתוך כתבי יד עבריים המתוארכים מתקופת ימי הביניים (https://sfardata.nli.org.il). המאגר צמח מתוך "מפעל הפליאוגרפיה העברית", שנוסד בשנת 1965 ביוזמתם של מלאכי בית אריה וקולט סיראט בחסות האקדמיה הלאומית למדעים בירושלים ובשיתוף פעולה עם המכון למחקר והיסטוריה של הטקסטים בפריז. עם התקדמות הטכנולוגיה, עבר המפעל שדרוגים מהותיים, והיה למערכת מקוונת המספקת גישה ציבורית וחיפושים מורכבים במידע הקודיקולוגי שנאסף.[7]

בניית המאגר כללה איתור כתבי יד עבריים הכוללים תאריך העתקה מפורש או אזכורים של שמות מעתיקים ותיעוד של כלל התכונות החומריות הנראות לעין – תכונות קודיקולוגיות – כמו מצע הכתיבה, סוג הדיו, הרכב הקונטרסים, עיטורים ועוד. כל מאפיין סווג והוגדר לפי קטגוריות אחידות על מנת לאפשר ניתוחים השוואתיים ויצירת טיפולוגיות קודיקולוגיות רחבות. מאגר ספרדתא משמש תשתית חשובה למחקר קודיקולוגי של הספר העברי, ומאפשר לחוקרים לאתר דפוסים היסטוריים, לזהות מגמות עיצוב והפקה ולהציע תיארוך והשוואות בין כתבי יד. בפועל כלי החיפוש של ספרדתא מאפשר למשתמשים לאתר כתבי יד מתוארכים בעלי מאפיינים דומים לאלו של כתב יד שאינו נושא תאריך, ובכך להציע תיארוך משוער

5 Colette Sirat, *Hebrew Manuscripts of the Middle Ages* (Cambridge: Cambridge University Press, 2002).

6 מלאכי בית אריה, **קודיקולוגיה עברית טיפולוגיה של מלאכת הספר העברי ועיצובו בימי-הביניים בהיבט היסטורי והשוואתי מתוך גישה כמותית המיוסדת על תיעוד כתבי-היד בציוני תאריך עד שנת 1540** (ירושלים והמבורג: האקדמיה הלאומית למדעים, 2021).

7 מלאכי בית אריה, **קודיקולוגיה עברית**, עמ' 34–35.

על בסיס דמיון טיפולוגי. עם זאת, תהליך זה דורש ידע קודיקולוגי מעמיק ויכולת ניסוח של שאילתות מדויקות בממשק החיפוש.

המחקר הנוכחי מבקש לבנות מודל שמסוגל לעבד את המידע הקודיקולוגי שנמצא בכתב היד ולהציע תיארוך משוער גם למשתמשים שאינם מומחים בתחום. כך יהיה ניתן להנגיש את ספרדתא לקהל רחב יותר ולשפר את הפונקציונליות המחקרית של הפרויקט כולו.

תיארוך של כתבי יד מבוסס נתונים קודיקולוגיים

קודיקולוגיה היא תחום מחקר העוסק בספר בצורת קודקס ובוחן אותו מכלל היבטיו הפיזיים, החומריים והתרבותיים: מן המבנה הטכני וחומרי ההפקה ועד להקשרים האינטלקטואליים והחברתיים שבהם הספר נוצר ונעשה בו שימוש. היבטים אלו חשובים בייחוד כאשר חסר מידע מפורש על אודות תאריך כתיבתו של המסמך.[8] את המונח "קודיקולוגיה" טבע פרנסואה מאזה (Mazé) בשנת 1950 והוא שיקף את התפתחותו של תחום חדש החוקר את הקודקס לא רק ככלי להעברת טקסט, אלא גם עדות חומרית מורכבת שיש לה ערך תרבותי, היסטורי וחברתי.[9]

דוגמה לחשיבותם של מאפיינים קודיקולוגיים לצורך תיארוך ניתן למצוא במחקרו של דניס נוסניטסין מאוניברסיטת המבורג. נוסניטסין בחן את האפשרות לתארך כתבי יד אתיופיים על בסיס צורת הניקוב והפיסוק (הכוונה לשיטות שרטוט השורות והכנת הדפים לכתיבה). הוא הציע חלוקה של כתבי היד לארבעה דפוסים עיקריים על פי מאפיינים טכניים אלו, והראה כי ניתן לזהות זיקה מובהקת בין הדפוסים לבין תקופות כתיבה שונות. תדירות הופעתם של דפוסים מסוימים בתקופות שונות משמשת, אפוא, כעוגן קודיקולוגי לתיארוך כתבי היד.[10]

דוגמה בולטת לתיארוך המבוסס על נתונים קודיקולוגיים ופליאוגרפיים היא מחקרה של עדנה אנגל, אשר בחנה קטעי כריכה (fragments) של כתב יד של מסכת מן התלמוד הבבלי. אנגל הצליחה לשער כי מקורו של כתב היד הוא באזור גרמניה או צרפת של ימינו, וכי הוא נכתב בין השנים 1270 ל-1400. בין המאפיינים הקודיקולוגיים שתמכו במסקנה זו ניתן למנות את שיטת עיבוד הקלף: הקלף היה חלק משני צדדיו – הן מצד הבשר והן מצד השיער – שיטה שהייתה נפוצה בגרמניה בלבד עד שלהי המאה ה-13. לכך נוספו מאפיינים נוספים

8 Albert Derolez, *The Palaeography of Gothic Manuscript Books: From the Twelfth to the Early Sixteenth Century* (Cambridge: Cambridge University Press, 2003); Michelle P. Brown, *Understanding Illuminated Manuscripts: A Guide to Technical Terms* (Los Angeles: The J. Paul Getty Museum, 2018).

9 בית אריה, **קודיקולוגיה עברית**, עמ' 47.

10 Denis Nosnitsin, "Pricking and Ruling in Ethiopic Manuscripts. An Aid for Dating?", *COMSt Bulletin*, 1/2 (2015): 94–109.

כגון תבנית הניקוב ומתווה פריסת הטקסט בעמוד.[11] דוגמה נוספת לשימוש במאפיינים קודיקולוגיים לצורך תיארוך מדויק היא מחקרו של פאבל גנקרצ׳יק (Gancarczyk) על הקודקס של סטראחוב – אסופה של יצירות מוזיקליות מלוות בטקסטים לטיניים, שתוארכה במקור לטווח שבין השנים 1460–1480. באמצעות ניתוח של פריסת הקונטרסים וסימני מים (watermarks)הצליח גנקרצ׳יק לצמצם מאוד את טווח התיארוך – מ-20 שנה לשלוש בלבד.[12] מחקרים אלה ממחישים את התרומה הפוטנציאלית של ניתוח קודיקולוגי מדוקדק לתיארוך כתבי יד. בהתאם לכך, המחקר הנוכחי מתמקד בתיארוך כתבי יד עבריים המסתמך על מאפיינים קודיקולוגיים כשדה מידע מרכזי.

אלגוריתמים לביצוע תיארוך כתבי יד

למידת מכונה (Machine Learning) היא תחום במדעי המחשב העוסק בפיתוח אלגוריתמים המסוגלים ללמוד מתוך דוגמאות ולהסיק מהן תחזיות או החלטות. ניתן לחלק את שיטות הלמידה לשני סוגים עיקריים: למידה מונחית (supervised learning) ולמידה בלתי מונחית (unsupervised learning). בלמידה מונחית האלגוריתם מאומן על בסיס אוסף של דוגמאות מתויגות כלומר נתונים אשר ידוע מראש מהי התוצאה הרצויה עבורם (למשל, כתבי יד שתאריך כתיבתם ידוע). האלגוריתם לומד מתוך הדוגמאות הללו כיצד לחזות את התוצאה עבור דוגמאות חדשות. בלמידה בלתי מונחית, לעומת זאת, אין תיוג מוקדם, והאלגוריתם מנסה לזהות דפוסים סמויים בעצמו מתוך הנתונים.[13]

במסגרת למידה מונחית, קיימות שתי שיטות עיקריות הרלוונטיות לתיארוך כתבי יד:

- סיווג (Classification) שבו המטרה היא לשייך כל כתב יד לקטגוריה מוגדרת מראש (למשל, לאיזו מאה הוא שייך).
- רגרסיה (Regression) שבה המטרה היא להעריך ערך מספרי מדויק, כמו שנת ההעתקה של כתב היד.[14]

11 Edna Engel, "A Codicological and Paleographical Analysis of the Sabbateni Hebrew Binding Fragments: Bavli Temurah Chapter 1", in Matthew S. Goldstone et al. (eds.), *Binding Fragments of Tractate Temurah and the Problem of Lishana Aḥarina*. (Leiden; Boston: Brill, 2018), pp. 40–53.

12 Paweł Gancarczyk, "The Dating and Chronology of the Strahov Codex", *Hudební věda* 2.43 (2006): 135–146.

13 Sathya Ramadass and Annamma Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification", *International Journal of Advanced Research in Artificial Intelligence* 2.2 (2013): 34–38.

14 Maruf A. Dhali, Camilo Nathan Jansen, Jan Willem de Witand and Lambert Schomaker, "Feature-Extraction Methods for Historical Manuscript Dating based on Writing Style Development", *Pattern Recognition Letters* 131 (2020): 413–420.

במחקר זה פיתחנו מודל דו-שלבי המשלב את שתי הגישות: בשלב הראשון מופעל אלגוריתם סיווג הקובע את המאה שבה נכתב כתב היד; ובשלב השני מופעל אלגוריתם רגרסיה, המעריך את שנת הכתיבה המדויקת בתוך אותה מאה. שילוב זה מאפשר לנצל את היתרונות של שתי הגישות – יציבות יחסית בזיהוי התקופה הכללית ודיוק גבוה יותר באומדן השנה.

תיארוך של כתבי יד באמצעות למידת מכונה

מרבית הגישות האוטומטיות לתיארוך כתבי יד מימי הביניים נשענות כיום על ניתוחים בלשניים או פליאוגרפיים של כתב היד.[15] בגישות מבוססות בלשנות משתמשים באלגוריתמים מתחום העיבוד של השפה הטבעית (NLP) אשר "מתרגמים" את המילים שבטקסט לייצוגים מספריים (וקטורים), ולאחר מכן מסווגים את כתב היד לטווח זמן מתאים על סמך מאפיינים לשוניים כמו תדירות מילים, מיקומן בטקסט והקשרן הייחודי לקורפוס מסוים.[16] ניתן גם לשלב בין ניתוח בלשני לניתוח פליאוגרפי הבוחן את סגנון הכתב ואת צורת האותיות.

דוגמה לשילוב כזה ניתן למצוא במחקר על כתבי יד שוודיים מימי הביניים, שבו תיארכו באמצעות שילוב ניתוח פליאוגרפי אוטומטי שהופק מתמונות דיגיטליות של כתבי היד עם מאפיינים לשוניים שנבחנו באמצעות אלגוריתמים של עיבוד שפה טבעית. תוצאות המחקר היו מרשימות: על פני אוסף רחב של כתבי יד שהתפרסו על פני כ-500 שנה, הטעות החציונית של המודל הייתה כ-12 שנים בלבד.[17] אתגר מהותי בגישות מסוג זה הוא בחירת המאפיינים (features) הרלוונטיים מתוך כתב היד. תהליך זה מצריך לא רק התערבות אנושית, אלא לעיתים מומחיות בתחום הקודיקולוגיה, איכות סריקה גבוהה או מדידות מדויקות.[18]

15 Fredrik Wahlberg, Lasse Mårtensson, and Anders Brun, "Large Scale Style-Based Dating of Medieval Manuscripts", *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing* (2015): 107–114.

16 Sidsel Boldsen, and Fredrik Wahlberg, "Survey and Reproduction of Computational Approaches to Dating of Historical Texts." *Nordic Conference on Computational Linguistics (NoDaLiDa)*, (Sweden: Linköping University Electronic Press, Sweden, 2021), pp. 145–156.

17 Fredrik Wahlberg, Lasse Mårtensson, and Anders Brun, "Large Scale Continuous Dating of Medieval Scribes Using a Combined Image and Language Model", *2016 12th IAPR Workshop on Document Analysis Systems (DAS)* (IEEE, 2016), pp. pp. 48–53.

18 Anmol Hamid, Maryam Bibi, Momina Moetesum and Imran Siddiqi, "Deep Learning Based Approach for Historical Manuscript Dating," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, (NSW, Australia: Sydney, 2019), pp. 967–972, doi: 10.1109/ICDAR.2019.00159.

שיטה רווחת נוספת היא ניתוח סגנון הכתב. כלומר זיהוי של תאריכים משוערים על בסיס שינויי צורה באותיות לאורך התקופות. מחקרים רבים מסתמכים על סריקות של כתבי יד, מזהים את האותיות באמצעות זיהוי תווים אופטי (OCR) ומאפיינים את סגנון הכתב באמצעות כלים של ראייה ממוחשבת.[19] כך לדוגמה, מחקר הולנדי שהתמקד במגילות ים המלח הצליח לתארך כתבים על בסיס ניתוח סגנון האותיות. 20 דוגמאות נוספות כוללות ניתוח של כתבי יד הולנדיים מן המאות ה-14–18 ומחקר רחב היקף על אלפי סריקות של כתבי-יד שוודיים – שניהם השתמשו בלמידת מכונה מבוססת פליאוגרפיה.[21]

המחקר הנוכחי מציע גישה שונה: שימוש במידע קודיקולוגי כבסיס ללמידת מכונה. כלומר לא נתונים פליאוגרפיים או לשוניים, אלא מאפיינים פיזיים-חומריים של כתב היד כמו סוג המצע, שיטת הניקוב, צבעי הדיו, מבנה העמוד וכדומה כפי שהם מתועדים על ידי קודיקולוגים. יתרונה של גישה זו בכך שמרבית המאפיינים ניתנים לזיהוי באמצעות התבוננות ישירה או מישוש של כתב היד גם ללא סריקה ברזולוציה גבוהה או ניתוח דיגיטלי מורכב.

גישה זו מאפשרת פיתוח ממשק שבו חוקר – גם אם אינו מתכנת או פליאוגרף – יוכל להזין בעצמו את המידע הקודיקולוגי מתוך צפייה בכתב היד הפיזי או הסרוק, ולקבל תיארוך משוער המבוסס על מודל למידת מכונה. בכך, נעשית הטכנולוגיה נגישה יותר לקהל רחב של חוקרי כתבי יד.

צמצום ובחירה של מאפיינים

בלמידת מכונה תהליך הבחירה והצמצום של מאפיינים (features), כלומר הנתונים שעל פיהם מתבצע החיזוי, הוא שלב מרכזי בתכנון המודל. לצמצום כזה יש יתרונות רבים: מקצר את זמן העיבוד, עשוי לשפר את דיוק התחזיות ומסייע להבין אילו נתונים משפיעים בפועל על התוצאה ובאיזה אופן.[22]

19 Maruf A. Dhali, Sheng He, Mladen Popović, Eibert Tigchelaar and Lambert, "A Digital Palaeographic Approach Towards Writer Identification in the Dead Sea Scrolls", in *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods: Volume 1: ICPRAM* (Portugal, 2017), pp. 693–702 https://doi.org/10.5220/0006249706930702

20 Dhali et al, "Feature-Extraction Methods for Historical Manuscript Dating Based on Writing Style Development".

21 He, Sammara, Burgers and Schomaker, "Towards Style-Based Dating", pp. 265–270; Wahlberg, Mårtensson and Brun, "Large Scale Style-Based".

22 Girish Chandrashekar and Ferat Sahin, "A Survey on Feature Selection Methods", *Computers & Electrical Engineering* 40.1 (2014): 16–28.

במחקר הנוכחי המניע העיקרי לצמצום המאפיינים היה יישומי: מטרתנו הייתה לפתח מודל שיוכל לתארך כתבי יד גם אם ברשות החוקר מצוי מידע חלקי בלבד. לשם כך, נדרשנו לזהות את אותם מאפיינים קודיקולוגיים שמנבאים את התאריך בצורה הטובה ביותר גם אם אינם מלווים במידע נוסף.

תהליך בחירת המאפיינים כלל שילוב של שתי גישות: ראשית, סינון אוטומטי באמצעות מודל ערכי SHAP – שיטה סטטיסטית המודדת את תרומת כל מאפיין לתחזית הסופית של המודל;[23] ושנית, בחירה ידנית מושכלת המבוססת על שיקולים של זמינות הנתונים, קלות הזיהוי שלהם והידע הקיים בקרב חוקרי כתבי יד. שילוב זה איפשר לנו ליצור מערך ממוקד של מאפיינים – כזה שמצד אחד מספק תיארוך מדויק ומצד שני מקל על החוקר בהפקת הנתונים הדרושים.

מטרות המחקר

1. לפתח מודל לחיזוי תאריכי העתקה של כתבי יד עבריים מימי הביניים באמצעות למידת מכונה מונחית, הנשענת על מאפיינים קודיקולוגיים.
2. להבין מהי תרומתו של כל מאפיין קודיקולוגי לתהליך התיארוך ואת אופן השפעתו על תחזית שנת הכתיבה.

שאלות המחקר

3. מהם המאפיינים הקודיקולוגיים החשובים ביותר לתיארוך כתבי יד באמצעות למידת מכונה על בסיס הנתונים הקיימים?
4. מהו מספר המאפיינים המינימלי הנדרש להשגת תיארוך מדויק באופן שיאפשר לחוקר להזין מידע מצומצם ככל האפשר?

מתודולוגיה

יצירת מאגר המידע

בשלב הראשון של המחקר נבנה מאגר מידע הכולל אלפי רשומות קודיקולוגיות מתוך אתר ספרדתא. הנתונים נאספו בעזרת גישה לתשתית הטכנית של האתר, המאפשרת שליפה שיטתית של המידע המוצג בו. כל רשומה נמשכה בנפרד, והנתונים שנאספו נשמרו לצורך עיבוד והכנה למודל.

[23] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee, "Consistent Individualized Feature Attribution for Tree Ensembles", *arXiv preprint arXiv:1802.03888* (2018).

הרשומות כללו מידע קודיקולוגי מגוון כגון סוג חומר הכתיבה, עימוד, שיטת הניקוב וצבעי הדיו ועוד, וכן מידע ביבליוגרפי כמו נושא החיבור, השפה ואזור ההפקה. מאחר שמרבית הנתונים הופיעו כטקסט חופשי או בקטגוריות מילוליות, היה צורך לעבדם בהמשך לצורה שניתנת לעיבוד חישובי.

המרה של מאגר המידע המילולי למבנה מספרי

המידע שנדלה מתוך אתר ספרדתא מופיע במקורו כטקסט מילולי, למשל, תיאור של חומר הכתיבה, שיטת הכריכה או אזור ההפקה. כדי שהנתונים יוכלו לשמש בסיס לאלגוריתמים של למידת מכונה, יש להמיר אותם לערכים מספריים. תהליך זה נעשה באמצעות שיטה הקרויה "משתנים מציינים" (dummy variables).

"משתנה מציין" הוא משתנה בינארי (0 או 1) שמסמן האם מאפיין מסוים קיים או לא קיים ברשומה. לדוגמה, אם באחת מהרשומות צוין כי "קונטרס מתחיל בצד – בשר", תיווצר עמודה בשם זה וכתב היד הרלוונטי יקבל בה את הערך 1 (כלומר: כן). אם כתב יד אחר אינו כולל תכונה זו, הערך יהיה 0. באופן זה נוצרת טבלה מספרית, שבה כל שורה מייצגת כתב יד וכל עמודה מציינת את נוכחותם או היעדרם של מאפיינים קודיקולוגיים שונים.

לצורך בניית רשימת המאפיינים שמהם הופקו המשתנים המציינים, נעשה שימוש במידע הזמין בממשק "החיפוש הסטטיסטי" של אתר ספרדתא. ממשק זה מאפשר לבצע שאילתות לפי חתכים קודיקולוגיים, למשל, חיפוש כל היחידות הקודיקולוגיות שהופקו באזור אשכנז, והוא כולל את כל סוגי המאפיינים הקיימים ברשומות המאגר. מידע זה שימש כבסיס למיפוי שיטתי של כלל המאפיינים האפשריים, אשר שימשו אותנו ליצירת מאגר הנתונים המספרי.

סינון מאגר המידע ובחירת מאפיינים

כדי לאמן מודל יעיל של למידת מכונה לתיארוך כתבי יד, יש לבחור בקפידה רק את אותם מאפיינים שתורמים בפועל לחיזוי התאריך. תהליך הסינון כלל שני שיקולים מרכזיים:

1. תרומתו של כל מאפיין למודל – עד כמה הנתון מסייע לחיזוי התאריך.
2. קלות הזיהוי והזמינות של הנתון – כדי לפתח מודל שמסוגל לפעול גם כאשר רק מידע חלקי נגיש לחוקר.

בשלב הראשון סוננו המאפיינים אוטומטית בהתבסס על ערכי SHAP – שיטה שנדונה לעיל ושמדרגת את תרומתו היחסית של כל מאפיין לתוצאה הסופית. מאפיינים שנמצאו כבעלי תרומה שולית הוסרו.

בשלב השני נערך סינון נוסף, שהתמקד ברמת הקושי שבזיהוי המאפיין מתוך כתב היד. מאחר שמאפיינים קודיקולוגיים רבים דורשים גישה פיזית לכתב היד או סריקה איכותית מאוד, ביקשנו לבדוק מהו המאמץ הנדרש בפועל מחוקר כדי לדלות כל אחד מהם. לשם כך,

הפצנו ״שאלון קושי בדליית נתונים קודיקולוגיים״ בקרב חוקרים בעלי מומחיות בקודיקולוגיה ובפליאוגרפיה עברית.

מן השאלון עלתה תמונה מורכבת: גם מאפיינים בעלי אופי פיזי-טכני – כמו צבע הדיו או שיטת השרטוט בחריטה – דורגו כקשים לזיהוי. הסיבה לכך היא שלרוב נדרש להחזיק את כתב היד עצמו או להשתמש בסריקות צבעוניות ברזולוציה גבוהה, שאינן תמיד זמינות. לעומת זאת, מאפיינים כמו לשון הטקסט (מעבר לעברית) או נושאו הכללי נחשבים לקלים יחסית לזיהוי, אך גם בהם ייתכנו טעויות, למשל, כאשר כתב היד אינו שלם או שקשה להבחין אם הוא כולל שפה נוספת.

טבלה 1 מציגה את ממוצע דרגות הקושי, לפי דיווחי המומחים (בסולם של 1 – קל מאוד, עד 5 – קשה מאוד):

טבלה מס׳ 1. תוצאות שאלון קושי בדליית נתונים קודיקולוגיים

סוג הנתון	ממוצע דרגת קושי
מצע הכתיבה	3.10
צבע הדיו	3.10
שרטוט בחריטה	2.70
נושא כללי של הטקסט	1.90
לשון הטקסט (מעבר לעברית)	1.90
אמצעים לשמירת סדר הקודקס	1.90
קישוטים	1.80
פורמט ומתווה הטקסט	1.70

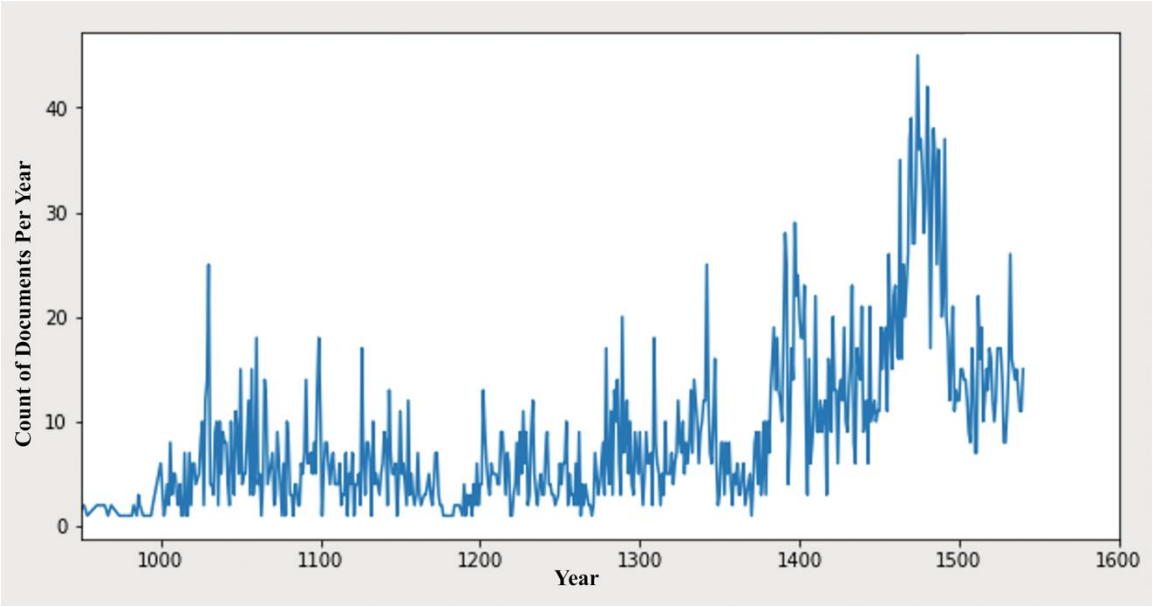
ממצאים

בסיס הנתונים

מאגר המידע ששימש למחקר נדלה מאתר ספרדתא. המאגר כולל 4,875 כתבי יד עבריים בעלי תאריך מדויק. כתב היד הקדום ביותר במאגר מתוארך לשנת 904, והמאוחר ביותר לשנת 1540. איור מספר 1 מציג את התפלגות כתבי היד על פני התקופות השונות.

מהתבוננות באיור עולה כי מרבית כתבי-היד המתוארכים מקורם במאות המאוחרות יותר. המאה ה 15 לבדה כוללת את מספר הרשומות הגבוה ביותר. התפלגות זו מצביעה על כך שהמידע הקיים אינו מאוזן, כלומר יש ריכוז גדול של נתונים מתקופה מסוימת לעומת מיעוט נתונים מתקופות אחרות. חוסר איזון כזה מקשה על המודל לבצע חיזוי מדויק עבור כתבי יד נדירים מהמאות המוקדמות יותר.

איור מס׳ 1. התפלגות כרונולוגית של כתבי היד המתוארכים. הציר האופקי מייצג את שנת יצירת היחידה הקודיקולוגית, והציר האנכי מציג את מספר כתבי היד שתוארכו לכל שנה.



מאגר המידע של ספרדתא כולל נתונים מקוטלגים לפי 14 קטגוריות שונות, שכל אחת מהן מייצגת סוג מסוים של מאפיינים קודיקולוגיים או ביבליוגרפיים. בכל קטגוריה יש מספר שונה של ערכים אפשריים. במונחי למידת מכונה, ניתן לראות בכל ערך כזה מאפיין בינארי – **(feature)** המציין האם תכונה מסוימת קיימת או נעדרת בכתב היד.

טבלה 2 מציגה את מספר המאפיינים בכל אחת מהקטגוריות. לדוגמה, בקטגוריה *ScriptString* הכוללת את סוגי הכתבים (כגון: אשכנזי, ספרדי, מזרחי וכדומה), יש ששה מאפיינים בלבד. לעומתה, הקטגוריה *MaterialString* העוסקת במצע הכתיבה יש 65 מאפיינים שונים.

מכיוון שלכל קטגוריה יש מספר שונה של מאפיינים, אי אפשר להשוות ביניהן לפי הכמות בלבד. כדי להבין איזו קטגוריה חשובה יותר לתיארוך, צריך לבדוק איזה חלק מהמאפיינים שלה נמצא שימושי, כלומר את האחוז מתוך כלל המאפיינים בכל קטגוריה שעבר את הסינון ונמצא רלוונטי.

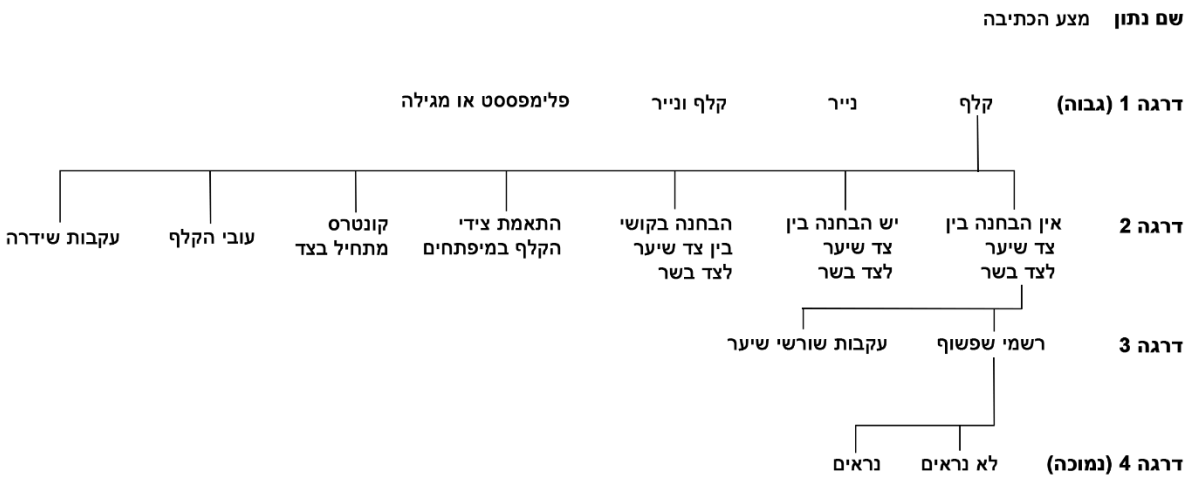
נוסף על כך, חשוב להבין כי מבנה המאפיינים בספרדתא הוא היררכי: כל מאפיין ראשי מחולק לתת מאפיינים, ולעיתים אף לדרגות עומק נוספות. כך למשל, המאפיין ״מצע כתיבה״ מתפצל לארבעה תתי מאפיינים: קלף, נייר, קלף ונייר, ופלימפססט או מגילה. כל אחד מהם

כולל תתי מאפיינים נוספים, עד לעומק של ארבע דרגות היררכיה. איור 2 מציג חלק מן ההיררכיה הזו, ודוגמה מלאה ניתן למצוא בממשק החיפוש של ספרדתא.

טבלה מס׳ 2. מאפייני המאגר המתוארך

שם עמודה	תאים מלאים	סוג ערכים	סוג נתון	סוג תוכן	מספר ערכים אפשרי
MaterialStringH	4335	מילולי	קודיקולוגי	מצע הכתיבה	65
RulingHardpointStringH	1970	מילולי	קודיקולוגי	שרטוט בחריטה	58
CatchwordsStringH	2946	מילולי	קודיקולוגי	אמצעים לשמירת סדר הקודקס	52
ErasStringH	3235	מילולי	קודיקולוגי	שיטות מניין השנים בקולופון	55
JustifiedLinesStringH	3202	מילולי	קודיקולוגי	עימוד השורה	35
IlluminationTransparencyStringH	3186	מילולי	קודיקולוגי	קישוטים	27
MitveStringH	3249	מילולי	קודיקולוגי	מיתווה	20
PrickingStringH	2520	מילולי	קודיקולוגי	ניקוב להדרכת שרטוט השורה	14
SubjectStringH	3661	מילולי	קודיקולוגי	נושא כללי של הטקסט	12
DestinationH	3722	מילולי	קודיקולוגי	יעד העתקה	9
InkStringH	3431	מילולי	קודיקולוגי	דיו	9
LanguageStringH	1102	מילולי	קודיקולוגי	לשון הטקסט חוץ מעברית	8
ScriptStringH	3542	מילולי	קודיקולוגי	כתב: טיפוס כתב	6
VocalizationStringH	1581	מילולי	קודיקולוגי	ניקוד הטקסט (כולל מסורה)	5

איור מס׳ 2. תרשים עץ חלקי – דרגות המאפיינים של הנתון הקודיקולוגי מצע הכתיבה

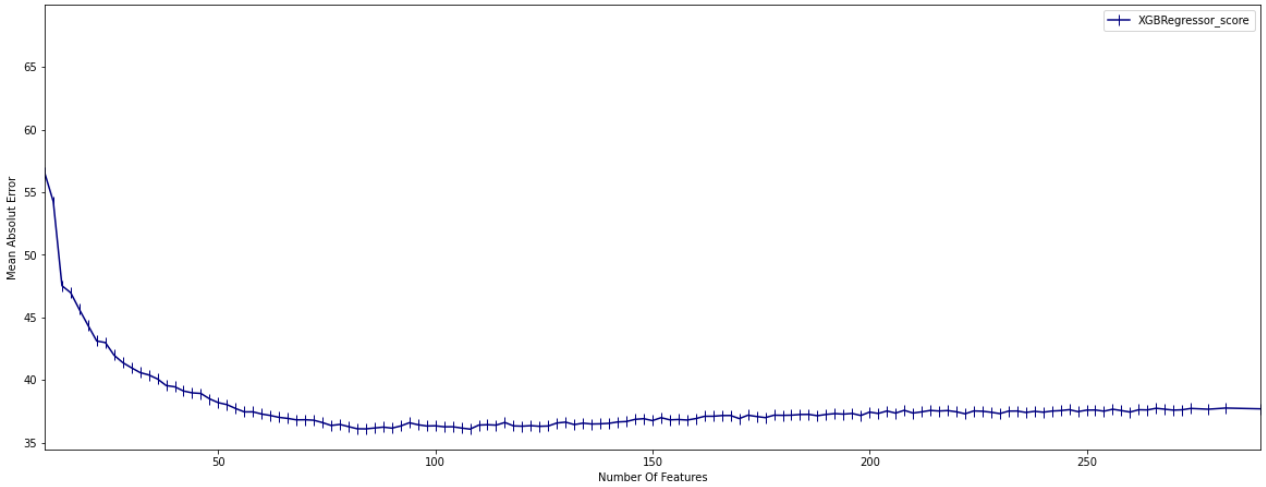


סינון הנתונים

מאגר הנתונים הגולמי כלל 375 מאפיינים שונים שנדלו מתוך אתר ספרדתא. מאחר שלא כל מאפיין תורם באותה מידה לתהליך החיזוי, אין הצדקה להשתמש בכולם. שימוש בעודף מאפיינים עלול להכביד על המודל, להאריך את זמן העיבוד ואף לפגוע בדיוק התוצאה. מטרת הסינון הייתה לזהות את כמות המאפיינים המינימלית הדרושה לשמירה על איכות חיזוי גבוהה, ולהבין מהם המאפיינים המשפיעים ביותר על תיארוך כתב היד.

לצורך הסינון, הוחלה שיטה סטטיסטית אוטומטית המבוססת על ערכי SHAP. מדובר בכלי שמדרג את התרומה של כל מאפיין לתחזית שהפיק המודל. איור 3 ממחיש את הקשר בין מספר המאפיינים לבין מידת הדיוק של התחזיות (לפי ממוצע הטעות האבסולוטית). מהגרף עולה שכאשר מספר המאפיינים יורד מתחת ל-66, חלה עלייה ברורה בשיעור השגיאות. על בסיס ממצא זה נקבע סף הסינון האוטומטי על 66 מאפיינים.

איור מס׳ 3. ממוצע הטעות כתלות במספר המאפיינים



בטבלה 3 מפורטת כמות המאפיינים בכל קטגוריה לפני תהליך הסינון האוטומטי ואחריו, וכן שיעור הצמצום (באחוזים):

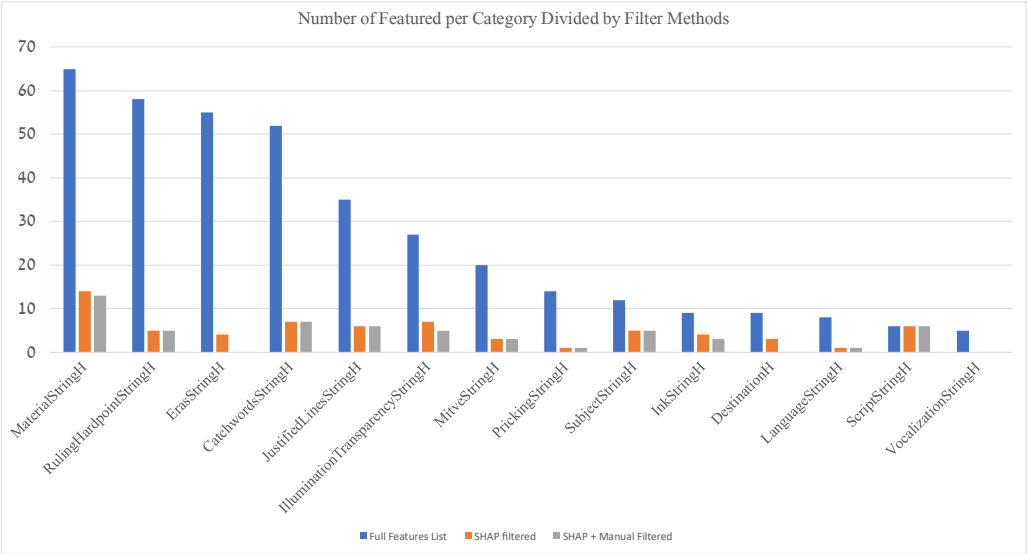
טבלה מס׳ 3. צמצום ובחירת מאפיינים באמצעות SHAP

שם עמודה במסד הנתונים המקורי	כמות מאפיינים לעמודה לפני סינון באמצעות SHAP	כמות מאפיינים אחרי SHAP	צמצום באחוזים
MaterialStringH	65	14	78%
RulingHardpointStringH	58	5	91%
ErasStringH	55	4	93%
CatchwordsStringH	52	7	87%
JustifiedLinesStringH	35	6	83%
IlluminationTransparencyStringH	27	7	74%
MitveStringH	20	3	85%
PrickingStringH	14	1	93%
SubjectStringH	12	5	58%
InkStringH	9	4	56%
DestinationH	9	3	67%
LanguageStringH	8	1	88%
ScriptStringH	6	6	0%
VocalizationStringH	5	0	100%

בשלב השני בוצע סינון ידני נוסף שהתבסס על ידע קודם ועל תוצאות שאלון המומחים (שתואר לעיל). מהשאלונים עלה כי מאפיינים מסוימים – כגון מידע מתוך הקולופון[24] או סוג הכתב – נחשבים קשים מאוד לדלייה, ולכן הוסרו מהמאגר. כך גם הוסרו קטגוריות כמו יעד ההעתקה ושיטת מניין השנים (שנשענות בעיקר על מידע מתוך הקולופון), וכן מאפיינים שאינם ניתנים לזיהוי בצילום, כמו עובי הקלף.

לאחר השלמת הסינון הידני, נותרו במאגר 53 מאפיינים בלבד. איור 4 מציג את כמות המאפיינים שנותרה בכל קטגוריה אחרי שני שלבי הסינון. ניתן לראות כי עיקר הצמצום התרחש כבר בשלב הסינון האוטומטי.

איור מס׳ 4. צמצום מאפיינים ובחירתם באמצעות SHAP וסינון ידני



24 ישנן שתי סיבות עיקריות אשר בגינן החלטנו לסנן נתונים קודיקולוגיים הקשורים לקולופון. הסיבה הראשונה היא היעדרותם של הנתונים הללו מרוב כתבי היד מהתקופה הזו. הסיבה השנייה ומהותית יותר היא שכאשר נתונים אלו קיימים, הם מייתרים את הרעיון של תיארוך באמצעות למידת מכונה מפני שלרוב הקולופון מציין את תאריך כתיבת המסמך.

חשיבות והשפעת המאפיינים על חיזוי תאריך כתב היד

במהלך המחקר נערכה בחינה שיטתית של תרומתו של כל מאפיין לתחזית שנת הכתיבה של כתב היד כדי להבין את עוצמת ההשפעה של המאפיינים השונים – הן הקודיקולוגיים והן אלו שמקורם בקטגוריות אחרות. חשוב לציין כי אף על פי שמוקד המחקר היה בנתונים קודיקולוגיים, המאגר כלל גם מאפיינים ביבליוגרפיים, טקסטואליים ופליאוגרפיים. מבין 13 הקטגוריות שנותרו לאחר הסינון האוטומטי, רק תשע עוסקות במאפיינים קודיקולוגיים. טבלה 4 מציגה את הקטגוריות שנותרו, סוג הנתונים שכל אחת מהן מייצגת ואת ציון ההשפעה האבסולוטי של כל קטגוריה לפי מדד SHAP המאפשר לזהות את עוצמת התרומה של כל מאפיין לתחזית המודל.

מן הנתונים עולה ממצא מעניין: שניים מתוך חמשת המאפיינים בעלי ההשפעה הגבוהה ביותר אינם קודיקולוגיים, אלא לקוחים מקטגוריות ביבליוגרפיות או טקסטואליות. כך, לדוגמה, הקטגוריה *יעד ההעתקה* הכוללת מידע על זהות המזמין שלמענו הועתק כתב היד דורגה במקום השני במדד ההשפעה של SHAP. ממצא זה מלמד כי שילוב בין סוגי נתונים שונים (ולא רק קודיקולוגיים) עשוי לתרום רבות לשיפור תיארוך כתבי היד באמצעות למידת מכונה.

טבלה מס׳ 4. עוצמת ההשפעה של הקטגוריות השונות על התיארוך (ציון SHAP)

שם העמודה בעברית	סוג נתונים	ציון השפעה אבסולוטי ממוצע
קישוטים / שקיפות הטקסט	מידע קודיקולוגי	1.42
יעד העתקה	מידע ביבליוגרפי	1.38
ניקוד הטקסט (כולל מסורה)	מידע קודיקולוגי	0.90
עימוד השורה	מידע קודיקולוגי	0.86
לשון הטקסט חוץ מעברית	מידע ביבליוגרפי	0.59
אמצעים לשמירת סדר הקודקס	מידע קודיקולוגי	0.55
דיו	מידע קודיקולוגי	0.53
הנושא הכללי של הטקסט	מידע ביבליוגרפי	0.53
מצע הכתיבה	מידע קודיקולוגי	0.49
שרטוט בעיפרון חורט	מידע קודיקולוגי	0.35
ניקוב להדרכת שרטוט השורה	מידע קודיקולוגי	0.18
כתב: טיפוס וסוג	מידע פליאוגרפי	0.11
שרטוט בחריטה	מידע קודיקולוגי	0.11

טבלה 5 מציגה את סכום השפעתם וממוצען של סוגי הנתונים השונים על חיזוי תאריך הכתיבה לפי מדד SHAP. מן הנתונים עולה כי המידע הקודיקולוגי תורם את החלק הארי לאיכות התיארוך (כ-74%). עם זאת, כאשר בוחנים את התרומה היחסית – כלומר את תרומתו של כל סוג מידע לעומת מספר המאפיינים שהוא כולל – הפער בין הקטגוריות מצטמצם, והתרומה של המידע הביבליוגרפי מתבררת כחשובה לא פחות. כך למשל, *יעד ההעתקה*, קטגוריה ביבליוגרפית, הוא אחד המאפיינים הבולטים בתרומתם לחיזוי.

טבלה מס' 5. עוצמת ההשפעה של סוגי הנתונים על תיארוך (ציון SHAP)

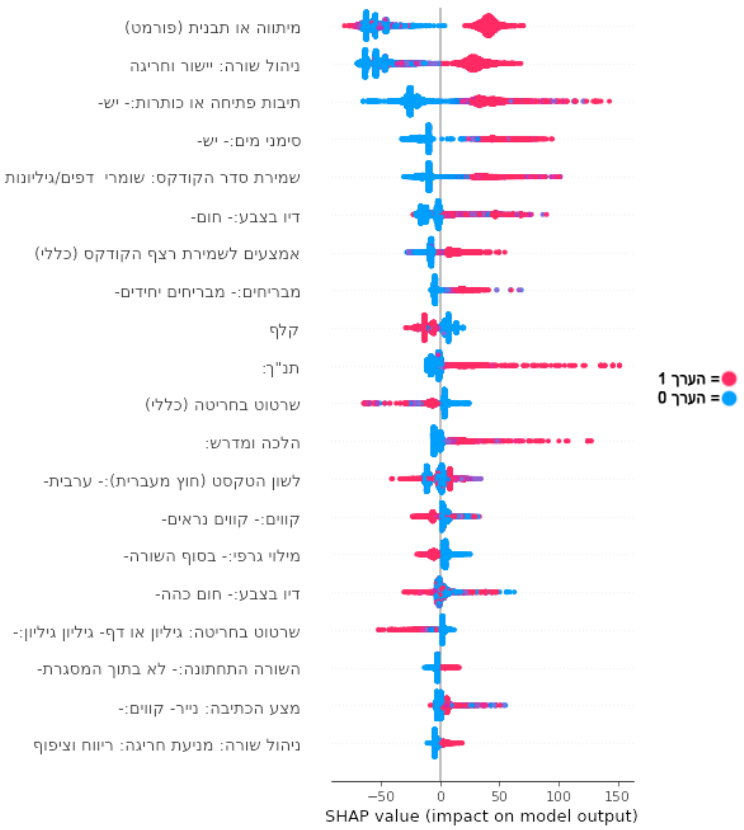
סוג נתונים	סכום ציוני השפעה אבסולוטית ממוצעת
מידע קודיקולוגי	5.92
מידע ביבליוגרפי	1.97
מידע פליאוגרפי	0.11
סה"כ	8

איור 5 מציג את 20 המאפיינים, שלפי מדדי SHAP, השפעתם על חיזוי שנת הכתיבה של כתבי היד היא הגבוהה ביותר. המדד מאפשר להעריך את התרומה של כל מאפיין לתחזית שהופקה עבור כל כתב יד.

כל שורה בגרף מייצגת מאפיין בודד, וכל נקודה מייצגת מופע של כתב יד שבו המאפיין קיים או נעדר. מיקומה של הנקודה בציר האופקי מציין את תרומתו של המאפיין לחיזוי – ערכים חיוביים מצביעים על תרומה לחיזוי של תאריך מאוחר יותר, וערכים שליליים מעידים על השפעה בכיוון של תיארוך מוקדם יותר.

הצבע של הנקודות מייצג את ערך המאפיין: אדום עבור ערך 1 (כלומר, המאפיין קיים) וכחול עבור ערך 0 (המאפיין לא קיים).

איור מס׳ 5. פיזור השפעתם של 20 המאפיינים המרכזיים על חיזוי שנת הכתיבה



מהאיור ניתן לזהות, למשל, כי נוכחות של כותרות פתיחה או תיבות (נקודות אדומות) מקושרת בעקביות עם חיזוי של תאריך כתיבה מאוחר יותר, ואילו היעדרן (נקודות כחולות) תורם לחיזוי מוקדם יותר. דפוס זה חוזר גם במאפיינים כמו ניהול שורות וסימני מים. חשוב לציין כי תרומתו של כל מאפיין תלויה בהקשרים נוספים, ולכן אותה תכונה עשויה להשפיע במידות שונות בהתאם לשילוב עם מאפיינים אחרים.

שונות בין תקופות: השפעת מאפיינים לפי מאה

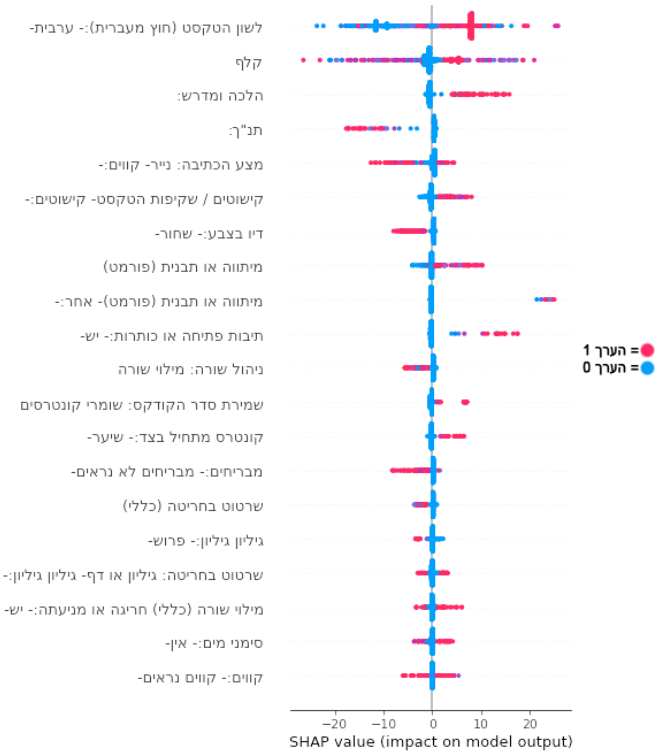
כדי להבין טוב יותר את ההשפעה המשתנה של מאפיינים על פני זמן, נבחנו דגמי SHAP נפרדים עבור כל אחת מהמאות ה-11 עד ה-16. באיורים 6א–6ו מוצגים 20 המאפיינים החשובים ביותר לחיזוי תאריך בכל מאה, כולל כיוון ההשפעה (חיובי/שלילי). לדוגמה:

- במאות ה-11 וה-12 מאפיין *לשון הטקסט (מעבר לעברית)* נמצא בין המשפיעים ביותר על תיארוך.

- במאות ה-15 וה-16 מאפיין זה כלל אינו מופיע ברשימה – עדות לשינוי בתפקידו ההיסטורי של מאפיין זה לאורך התקופות.

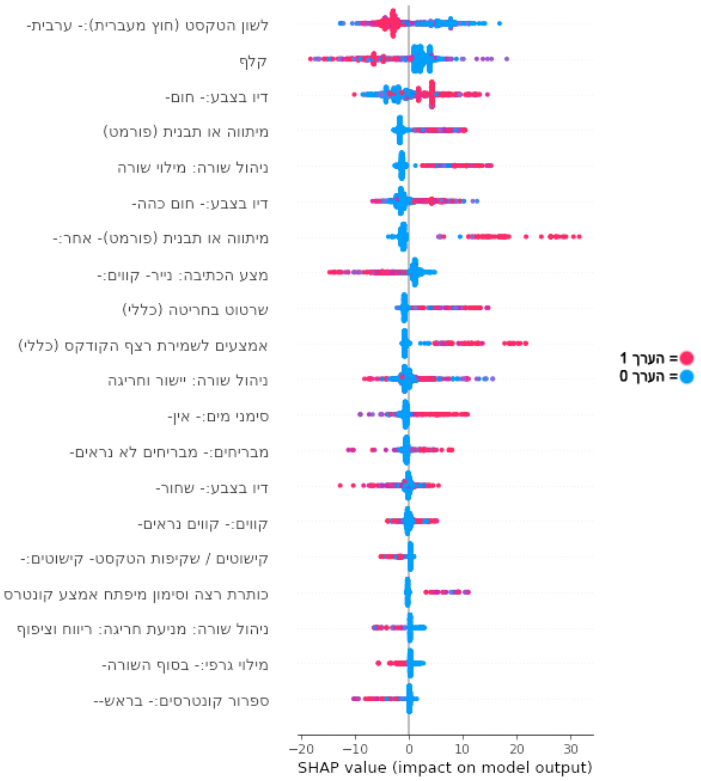
גם כיוון ההשפעה משתנה: מאפיין *קלף כמצע כתיבה* נוטה להפחית את השנה החזויה במאה ה-12, אך במאה ה-13 אותו מאפיין דווקא מעלה את התחזית. שינויים אלה מעידים על מעבר הדרגתי בדפוסי השימוש בקלף ובטכנולוגיות הכתיבה לאורך הדורות.

איור 6. א. השפעת מאפיינים לפי מאה – המאה ה-11

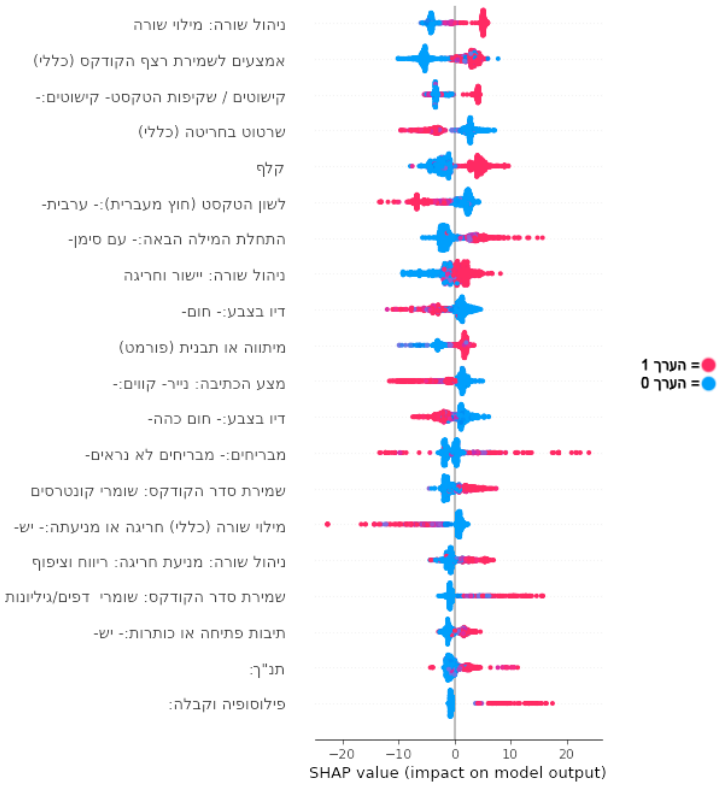


.

איור 6. ב. השפעת מאפיינים לפי מאה – המאה ה-12

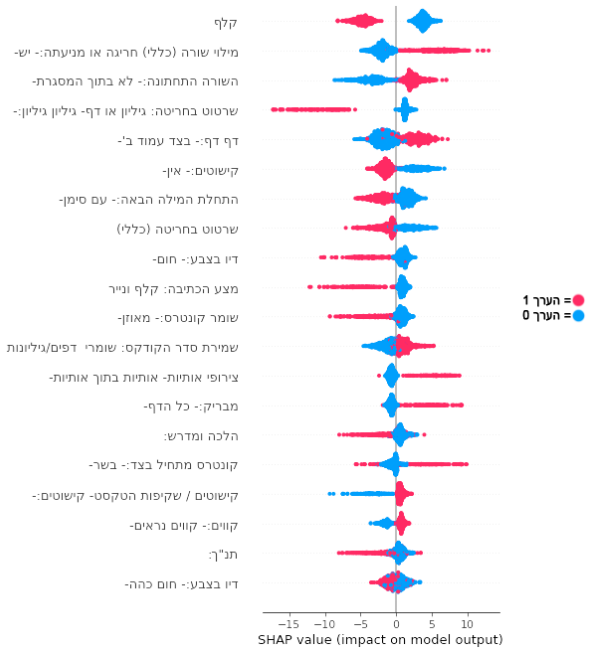


איור 6. ג. השפעת מאפיינים לפי מאה – המאה ה-13

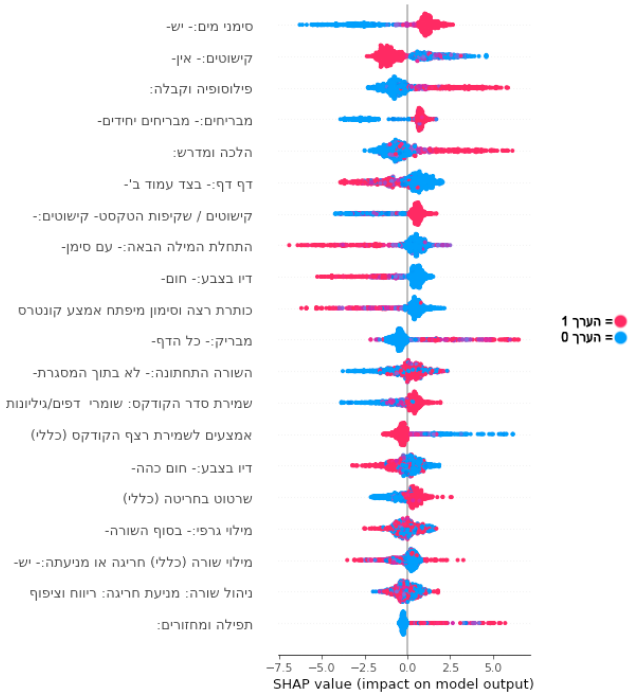


איור 6.ד. השפעת מאפיינים לפי מאה – המאה ה-14

איור 6.ה. השפעת מאפיינים לפי מאה – המאה ה-15



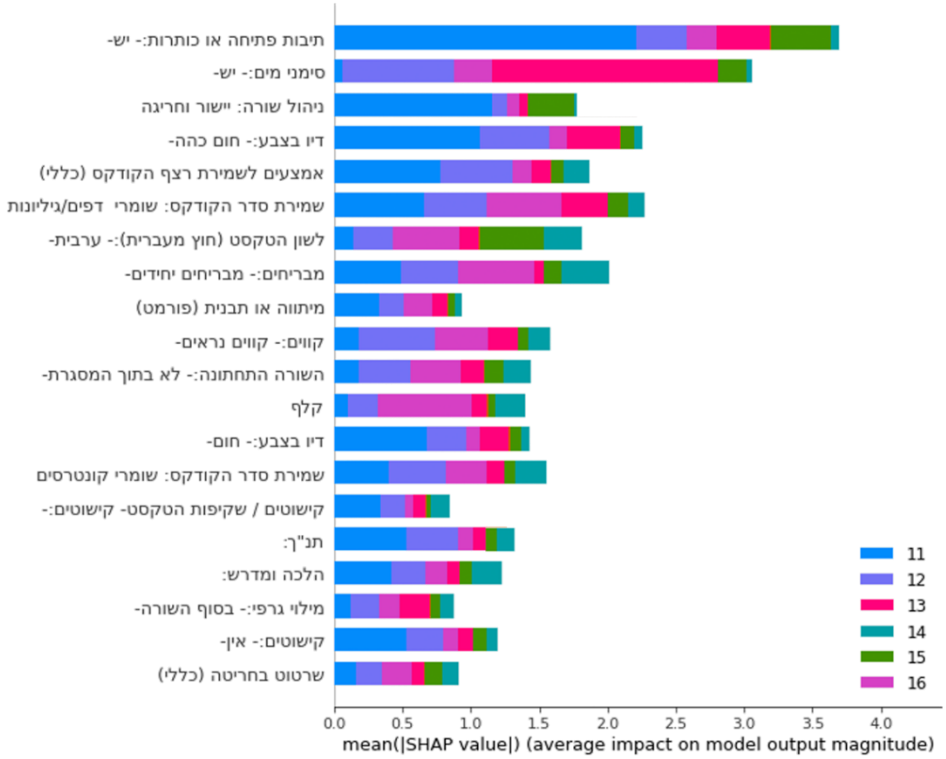
איור 6.ו. השפעת מאפיינים לפי מאה – המאה ה-16



מאיורים 6א–6ו עולה כי כאשר בוחנים את השפעת המאפיינים על תיארוך כתבי היד בכל מאה בנפרד, מתגלה שוני ניכר בדפוסי ההשפעה. כלומר המאפיינים שתורמים לתיארוך מדויק במאה מסוימת אינם בהכרח מובהקים במאות אחרות, ולעיתים אף מפעילים השפעה הפוכה.

איור 7 מסכם את ממצאי מודל SHAP עבור כל אחת מהמאה ה-11 עד ה-16, ומציג את 20 המאפיינים שהשפעתם על חיזוי המאה היא הגדולה ביותר. מהניתוח עולה, לדוגמה, כי למאפיין *תיבות פתיחה או כותרת* הייתה תרומה רבה לחיזוי במאה ה־11, ואילו בכתבי יד מאוחרים יותר השפעתו נחלשת באופן ניכר. ממצאים אלו מדגישים את חשיבות הבחינה ההקשרית של כל מאפיין – לא רק כשלעצמו, אלא ביחס לתקופה ההיסטורית שבה נוצר כתב היד.

איור מס׳ 7. השפעת מאפיינים קודיקולוגיים על חיזוי המאה שבה נוצרה היחידה הקודיקולוגית, לפי מודל SHAP



דיון

מחקר זה בחן את האפשרות לתארך כתבי יד עבריים מימי הביניים באמצעות מודל של למידת מכונה מונחית, שהתבסס על מאפיינים קודיקולוגיים שנדלו ממאגר המידע של ספרדתא. המטרה לא הייתה רק לבנות מנגנון חיזוי מדויק ככל האפשר, אלא גם להבין אילו מאפיינים תורמים בפועל לתהליך התיארוך, וכיצד ניתן לצמצם את כמות הנתונים הנדרשים, כך שהשיטה תהיה נגישה וישימה גם לחוקרים שאין להם רקע טכנולוגי.

שתי שאלות מרכזיות עמדו בבסיס המחקר: מהם המאפיינים הקודיקולוגיים החשובים ביותר לתיארוך של כתבי יד עבריים? ומהו מספר המאפיינים המינימלי הנדרש כדי להשיג תיארוך מדויק?

השאלה הראשונה נגעה לצמצום כמות הנתונים הדרושים לתיארוך. אף שבספרות מוזכרות לרוב שלוש סיבות עיקריות לכך – קיצור זמן החישוב, שיפור הדיוק והבנת החשיבות היחסית של כל מאפיין – המחקר הנוכחי מבקש להוסיף סיבה נוספת המשקפת את תנאי העבודה של חוקרי כתבי יד בפועל: הקלה על תהליך איסוף הנתונים, מתוך הבנה שלא כל מאפיין ניתן לזיהוי בקלות דרך סריקה או צילום של כתב היד.

תהליך הצמצום נעשה בשני שלבים. בשלב הראשון הופעל סינון אוטומטי באמצעות מדד SHAP ששדרג את תרומתו של כל מאפיין לתחזית הסופית. שלב זה הפחית בכ-80% את מספר המאפיינים, אך שמר על דיוק גבוה יחסית. בשלב השני בוצע סינון ידני שהתבסס על הערכת מומחים לגבי הקושי בזיהוי מאפיינים מסוימים מגרסה דיגיטלית. לשם כך גובש שאלון ייעודי, שנשלח לחוקרים בתחומים הרלוונטיים. בין המאפיינים שסוננו נכללו נתונים הנדלים מן הקולופון – שלעיתים כולל את תאריך ההעתקה באופן מפורש – וכן מאפיינים פליאוגרפיים מתקדמים כדוגמת סיווג טיפוס הכתב, הדורשים מומחיות גבוהה. גם הבחנה בין צד השיער וצד הבשר של קלף או בין סוגי נייר הוגדרה כקשה כאשר אין גישה פיזית למסמך. אף על פי שהסינון הידני הוביל לירידה קלה בדיוק, הוא שיפר את הישימות המחקרית של המודל בכך שהוא מצריך פחות מידע ראשוני ומאפשר עבודה המבוססת על דיגיטציה.

השאלה השנייה עסקה בזיהוי המאפיינים המשפיעים ביותר על תיארוך כתב היד. מודל SHAP אִפשר לא רק לדרג את עוצמת ההשפעה של כל מאפיין, אלא גם להבין את כיוונה, כלומר האם נוכחותו של מאפיין מסוים גורמת לחיזוי תאריך מוקדם יותר או מאוחר יותר, ואף לזהות שינויים בהשפעתם של מאפיינים אלה בין תקופות שונות. מן הממצאים עולה כי מאפיינים אינם פועלים במנותק מהקשרם. מאפיין שעשוי לשמש סימן מובהק לתיארוך במאה ה-11 עלול לאבד מחשיבותו או לשנות את השפעתו במאה ה-13. תובנה זו מדגישה את אופיים ההיסטורי והתרבותי של כתבי היד ואת הצורך בשיקול דעת פרשני גם כאשר נעזרים במודלים חישוביים.

בחינת ספרות המחקר מעלה כי רבים מהמחקרים שהתמקדו בתיארוך קודיקולוגי – כמו זה של דניס נוסניטסין על כתבי יד אתיופיים,[25] או עבודתה של עדנה אנגל על קטעים תלמודיים – הדגישו מאפיינים פיזיים מדרגות היררכיה נמוכות (איור 2), כגון פריסת קונטרסים או סימני מים.[26] מאפיינים אלה מצריכים לעיתים גישה ישירה לכתב היד, וכן מומחיות בזיהוי. גם בעבודתו של גנצ׳ארצ׳יק נעשה שימוש במאפיינים מסוג זה.[27]

לעומת זאת, במחקר הנוכחי עולה כי מרבית המאפיינים שהמודל זיהה כהכרחיים לתיארוך משתייכים דווקא לדרגות הגבוהות של ההיררכיה הקודיקולוגית. אלה מאפיינים כוללים יותר, שניתן לזהותם גם בתצפית ראשונית בכתב יד סרוק. כך למשל, עצם קיומו של מצע כתיבה מסוג קלף – מידע שעל פניו נראה ״טכני״ בלבד – טומן בחובו משמעות גאוגרפית. קלף היה בשימוש נרחב באשכנז וביזנטיון עד המאה ה-15, אך במזרח הוחלף בנייר כבר במאה ה-11.[28] אף שהמודל לא ״הוזן״ באופן ישיר באינפורמציה על אזור

25 Denis Nisnitsin, ״Pricking and Ruling״ pp. 94–109.

26 Edna Engel, "A Codicological and Paleographical", pp. 40–53.

27 Paweł Gancarczyk, "The Dating and Chronology of the Strahov Codex" *Hudební věda* 2.43 (2006): 135–146.

28 מלאכי בית אריה, **קודיקולוגיה עברית**.

ההעתקה, הוא הצליח לזהות את ההבדלים הללו דרך צירופים של מאפיינים עקיפים כמו מצע כתיבה, צבע דיו, וסוגי סימנים לשמירת סדר הקודקס.

מאפיינים נוספים שזוהו כמשפיעים במיוחד הם צבע הדיו ושיטת שמירת הסדר הפנימי: שומרי דפים לעומת שומרי גיליונות. ממצאינו עולים בקנה אחד עם הידוע מן הספרות: שומרי גיליונות הופיעו באירופה החל מהמאה ה-14, אך כבר כתב יד ספרדי משנת 1225 כולל שימוש בהם. לעומתם, שומרי דפים הופיעו לראשונה בכתב יד מזרחי מן המאה ה-12, והיו לנפוצים במאה ה-14 עם התפשטות הנייר. אף ששיטת שומרי הדפים נחשבת לפחות חסכונית, היא נפוצה יותר: כ-40% מהכתבים במאגר נכתבו בשיטה זו, לעומת 4% בלבד שעשו שימוש בשיטת שומרי הגיליונות.[29] גם כאן, כמו בדוגמאות קודמות, נדרשת לרוב ידיעה מוקדמת על אזור ההפקה של כתב היד כדי להיעזר במאפיינים אלו באופן ישיר בתיארוך. עם זאת, המודל המוצע במחקר זה מצליח לפעול ללא מידע כזה ולזהות את המאפיינים והצירופים שמאפיינים תקופות ואזורים – גם אם המשתמש עצמו אינו מזין אותם במפורש.

אחד מיתרונות המודל שפותח במחקר זה הוא האפשרות להציג את תרומתם של מאפיינים שונים לתיארוך ולציין את התקופה שאליה משתייך כתב היד. כאשר לחוקר יש הערכה מוקדמת – גם אם חלקית – באשר לתקופה שבה הועתק כתב היד, הוא יכול להיעזר בתרשימים 6א–6ו כדי לזהות מהם המאפיינים הרלוונטיים ביותר לאותה מאה. בכך מתאפשר מיקוד יעיל בתהליך איסוף הנתונים וחיסכון בזמן ובמשאבים.

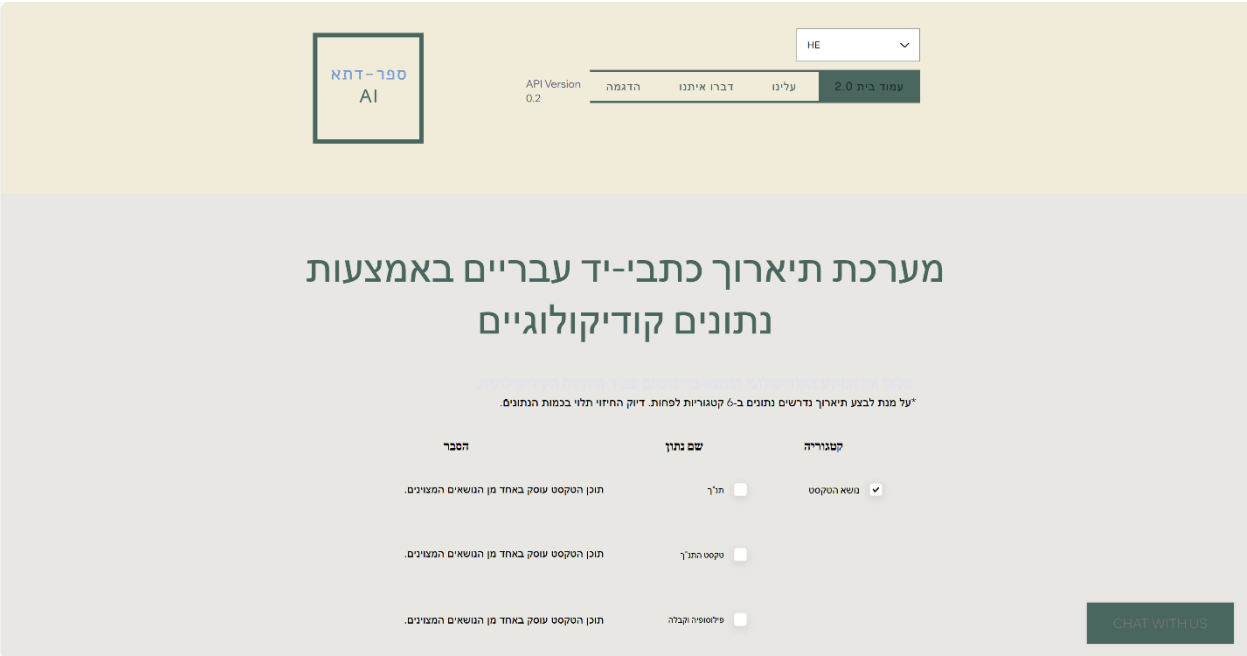
מעבר לשימוש האופרטיבי בתיארוך, הנתונים שנאספו במסגרת המחקר עשויים לתרום להבנה רחבה יותר של ההתפתחות ההיסטורית של הספר העברי. ניתוח השכיחות וההשפעה של מאפיינים מסוימים לאורך זמן מאפשר לשרטט מגמות תרבותיות וחומריות, ולעיתים אף לזהות מעבר בין מסורות שונות. כך נפתח פתח למחקר משווה ולגיבוש טיפולוגיות חדשות על בסיס עקרונות אמפיריים, המשלבות תובנות היסטוריות, קודיקולוגיות ודיגיטליות גם יחד.

דוגמה ליישום המודל המוצע

המודל שפותח במסגרת מחקר זה זמין לשימוש באמצעות ממשק אינטרנטי ניסיוני ספר-דתא AI (https://www.sfardata-ai.com) המאפשר לכל חוקר להזין מאפיינים קודיקולוגיים שנצפו בכתב יד סרוק ולקבל תיארוך משוער המבוסס על חישוב סטטיסטי (איור 8). המערכת נגישה לציבור החוקרים, ואינה דורשת מומחיות פליאוגרפית או ידע טכני מתקדם. היא מבוססת על מאגר מאפיינים שהוגדרו מראש והמתאימים לכתבי יד דיגיטליים. בכך נפתח פתח לשימוש מעשי במודל גם בקרב חוקרי מדעי הרוח, המעוניינים להיעזר בכלי חישובי לתיארוך ראשוני או לבחינה ביקורתית של תאריכים קיימים.

[29] מלאכי בית אריה, **קודיקולוגיה עברית**.

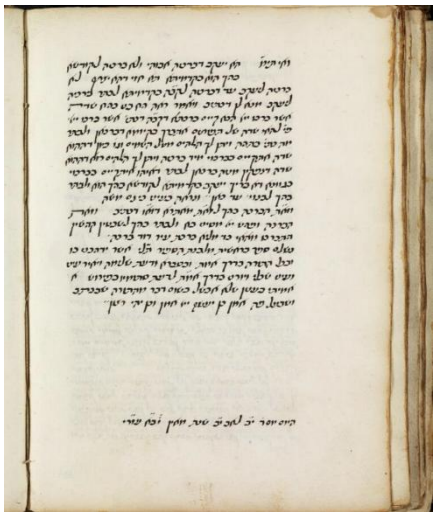
איור 8: המסך הראשי של מערכת ספר-דתא – AI ממשק אינטרנטי לתיארוך כתבי יד עבריים על סמך מאפיינים קודיקולוגיים



כדי לבחון את השימושיות המעשית של המודל שפיתחנו, נבדק מקרה של כתב יד שתוארך בתאריך מסוים במאגר ספרדתא – בהתבסס על הקולופון שבו – אך המודל שלנו הציע תאריך שונה. מדובר בכתב יד מהאוסף של הספרייה הלאומית המרכזית בפירנצה, איטליה, שסימונו Ms. Magl. XL.72 (מספר זיהוי בספרדתא: YY 175, מספר סרט בספרייה הלאומית: ס׳ 11987). כתב היד כולל את החיבור ״לבנת הספיר על ספר בראשית״, שהוא פירוש לקבלת הזוהר שכתב ר׳ יוסף אנג׳לט בסרגוסה בשנת פ״ה (1325).[30] בקולופון, שנמצא בדף 141ב, נכתב: ״היום יום ד׳ י״ב לאב י״ב שנת מאין יבא עזרי״ (את צילום הקולופון ניתן לראות באיור 9). אף על פי שנראה כי הקולופון מצביע על תאריך ברור, המודל שפיתחנו הציע תאריך שונה מהתאריך בספרדתא, והדבר עורר שאלות על מהימנות התאריך, פיענוחו או מקורות הטעות האפשריים. דוגמה זו ממחישה את הפוטנציאל של המודל לשמש כלי עזר לבחינה מחודשת של תיארוך כתבי יד, גם כאשר קיים תאריך לכאורה.

30 על המחבר והחיבור ראה: איריס פליקס, **פרקים בהגותו הקבלית של הרב יוסף אנג׳לט**, עבודת גמר לתואר שני, האוניברסיטה העברית בירושלים, תשנ״א, עמ׳ 1–6.

איור מספר 9: כתב יד הספרייה הלאומית המרכזית של פירנצה Ms. Magl. XL.72 דף 141ב'[31]



במאגר ספרדתא תוארך כתב היד לשנת 1527 בהתבסס על פרשנות מילולית וקלנדרית של המילים "יבא עזרי", שגימטרייתן היא רפ"ז. ההסבר לקביעת התאריך המשוער לשנת 1527 הוא רמז שנכתב בקולופון. כך נכתב ברשומת כתב היד בקטלוג הספרייה הלאומית: "היום יום ד' יב לאב י"ב שנת מאין יבא עזרי' המלים 'יבא עזרי' מסומנות, אך הפרט אינו יכול להיות שנת ש' [סכום האותיות] כיון שהיא אינה מתאימה ללוח, מעל האות י' מסומן סימן נוסף, ואולי הוא בא לבטל את המילה 'יבא' כולה. אם כן נשאר הפרט של עזרי=שנת רפ"ז שבה יב אב חל ביום ד בשבוע". אולם בבחינה מדוקדקת יותר של הקולופון עולה כי האותיות המסומנות הן דווקא: י', ב', א', ז', ר' כלומר י"ב אב שנת ר"כ (1460). שני הקווים שמעל האות יו"ד, אשר פורשו כסימן מחיקה, מופיעים גם במקומות אחרים בכתב היד ואינם בהכרח מעידים על ביטול המילה. גם מבחינה קלנדרית, י"ב באב אכן חל ביום ד' באותה שנה. ממצא זה קרוב מאוד לתחזית שהתקבלה מן המודל שפיתחנו (איור 10): שנת 1457 – מוקדמת בכ-70 שנה מהתאריך הרשום בספרדתא.

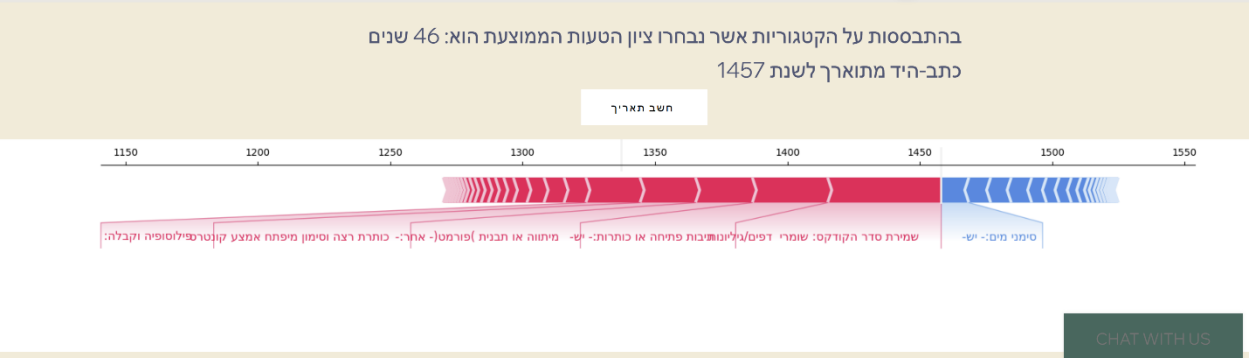
חיזוק נוסף לתאריך המוקדם יותר התקבל מכתב יד אחר של אותו מעתיק שמצוי בספריית אוניברסיטת בר-אילן. ד"ר יעקב פוקס ממחלקת כתבי יד בספרייה הלאומית זיהה עותק נוסף של החיבור "לבנת הספיר" אשר ככל הנראה נכתב בידי אותו מעתיק של כתב היד פירנצה, Ms. Magl. XL.72. בכתב היד המצוי בבר-אילן, הכתוב על נייר, מופיעים סימני מים ברורים (למשל בדפים 5, 6, 143): ראש שור עם פרח שיש חמישה עלי כותרת מעליו.

[31] From the collection of the Biblioteca Nazionale Central di Firenze, The National Library of Israel. "Ktiv" Project, The National Library of Israel. https://jewish-faculty.biu.ac.il/files/jewish-faculty/shared/JSIJ25/prebor.pdf

סימן זה זהה לדגם הידוע מתוך המאגר (*Wasserzeichen des Mittelalters*) WZMA[32] אשר מתוארך לשנת 1455 (ראו איור 11).

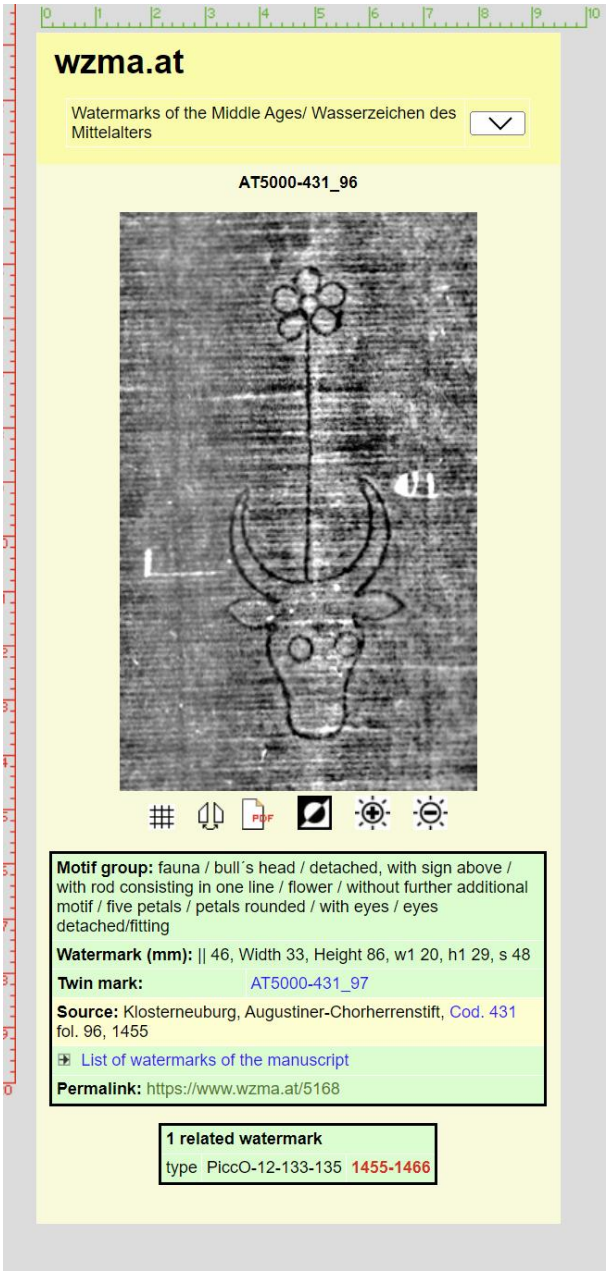
המקרה מדגים היטב את כוחו של המודל שאינו רק טמון ביכולתו להציע תיארוך עצמאי, אלא גם ביכולתו לשמש כלי ביקורתי ובוחן שמאתגר תיארוכים קיימים, מציף נקודות בעייתיות בפרשנות הקולופון ומכוון לבדיקה חוזרת של ממצאים באמצעות ראיות חיצוניות כגון סימני מים, טיפוס הכתב ופרטי הפקה חומריים.

איור מספר 10: תיארוך כתב יד הספרייה הלאומית המרכזית של פירנצה Ms. Magl. XL.72 במערכת https://sfardata-ai.com



32 https://www.wzma.at/infos.php. תודה רבה לד״ר יעקב פוקס על שהפנה אותי לכתב היד ולאתר הרלוונטי, וסייע לי בפיענוח סימן המים. תודתי גם לדודי בן נעים, אוצר הספרים הנדירים בספריית אוניברסיטת בר-אילן, על עזרתו באיתור הסימנים ובפיענוחם.

איור מספר 11: סימן מים בצורת ראש שור מתוך
Wasserzeichen des Mittelalters - wzma.at[33]



[33] Permalink: https://www.wzma.at/5168

סיכום

מחקר זה הציג שיטה לתיארוך כתבי-יד עבריים מימי הביניים המבוססת על למידת מכונה מונחית ועל שימוש בנתונים קודיקולוגיים ביבליוגרפיים וטקסטואליים. השיטה נשענת על מפעלם רב השנים של מלאכי בית-אריה וקולט סיראט ועל הנתונים שנדלו ממסד הנתונים ספרדתא.

המחקר עסק בפיתוח כלי לתיארוך אוטומטי ובהבנת ההשפעה של המאפיינים הקודיקולוגיים השונים על תוצאת התיארוך. המחקר שילב בין ידע מהספרות, תובנות ממומחים בתחום וממצאים של מודל SHAP. תהליך זה צמצם מאוד את כמות הנתונים הדרושה להפעלת המודל: רק כ־19% מהמאפיינים הקיימים באתר ספרדתא נדרשים בפועל לתיארוך מדויק.

התוצר המרכזי הוא כלי יישומי לחוקרי מדעי הרוח המתארך כתבי-יד עבריים על סמך מאפיינים קודיקולוגיים בלבד. כאשר החוקר מזין נתונים עבור כל 11 הקטגוריות, הטעות הממוצעת עומדת על כ-39 שנה. המערכת גמישה ומאפשרת גם תיארוך על בסיס הזנת נתונים חלקית – לפחות 6 קטגוריות מתוך ה-11 – לצד הצגת אומדן לטעות החזויה בהתאם לנתונים שהוזנו.

איכות התיארוך אינה תלויה רק בכמות הנתונים אלא גם באיכותם ובשילוביהם: מאפיינים שונים משפיעים זה על זה ומשנים את כיוון התחזית ועוצמתה בהתאם להקשרים. ממשק המערכת (בכתובת https://sfardata-ai.com) כולל תצוגה של הטעות החזויה בזמן אמת, בהתאם להרכב הנתונים.

מגבלות מחקר והצעות למחקרי המשך

המחקר מתבסס על 4,875 יחידות קודיקולוגיות מתוארכות המתפרסות בטווח המאות ה-11 עד ה-16 באופן שאינו מאוזן. פיזור זה מגביל את יכולת המודל לתארך כתבי-יד מתקופות אחרות, ופוגע בדיוק התיארוך בפרקי זמן שאינם מיוצגים היטב.

מגבלה נוספת היא חוסר נתונים ברשומות רבות. היעדר זה משפיע לא רק על ביצועי המודל, אלא גם על אופן פיענוח הנתונים: המרת ערכים מילוליים למשתנים בינאריים (״1״ או ״0״) יוצרת מצב שבו *אי קיום מידע* (לדוגמה, כאשר אין מידע על צבע הדיו) וגם *שלילה מפורשת* (הכתב אינו בדיו חום, לדוגמה) מקבלים את אותו ערך (״0״). בכך נוצר טשטוש בין ״לא ידוע״ ל״שלילה״ שעלול להטעות את האלגוריתם.

כדי להתגבר על מגבלות אלה, מומלץ להרחיב את בסיס הנתונים – הן את מספר כתבי-היד והן את טווח השנים. הרחבה כזו יכולה להיעשות באמצעות דלייה של מידע קודיקולוגי מתצלומים כולל יחידות קודיקולוגיות שעוד לא תויגו במסד הנתונים. בעתיד ניתן יהיה לשלב

אלגוריתמים של ראייה ממוחשבת כדי לזהות מאפיינים קודיקולוגיים אוטומטית, למשל, סוג הדיו, שיטת קיפול או טיפוס הכתב.

כבר כיום קיימת דוגמה ראשונית לכך: חוקרים הצליחו לאמן אלגוריתם לזיהוי טיפוסי כתב עבריים מתוך תצלומים של כתבי-יד.[34] המשך פיתוח בתחום זה יוכל להוביל לצמצום החוסר ברשומות קיימות ולהוספת כתבי-יד חדשים למסד הנתונים – ובכך לשפר את דיוק המודל ולהרחיב את טווח השימוש בו.

[34] Droby, Rabaev, Vasyutinsky Shapira, Kurar Barakat and El-Sana, "Digital Hebrew Paleography". https://jewish-faculty.biu.ac.il/files/jewish-faculty/shared/JSIJ25/prebor.pdf